A Projection-based Segmentation Algorithm for Breaking MSN and YAHOO CAPTCHAs

Shih-Yu Huang, Yeuan-Kuen Lee, Graeme Bell and Zhan-he Ou

Abstract—**Defeating a CAPTCHA test requires two procedures: segmentation and recognition. Recent research shows that the problem of segmentation is much harder than recognition. In this paper, a new projection-based segmentation algorithm is proposed for the MSN and Yahoo CAPTCHAs. Experimental results show that the proposed algorithm can improve correct segmentation rates ranging from 9% to 14% over the traditional one.**

Index Terms—**CAPTCHA, segmentation, recognition**

I. INTRODUCTION

As the internet increases in terms of size and in terms of available services, people gain more convenience but also face new challenges. Free services on the internet may be abused by automated computer programs (often referred to as scripts or bots – here, we use bot). Such bots may be intended to broadcast junk emails, post advertisements, or ask the server to respond at a very high frequency. All of these forms of misuse will decrease the usefulness of internet services. To prevent such abuses, it is very important to design an automatic system to differentiate between the messages of legitimate human users and non-legitimate computer bots. The Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) was created to address these needs [1].

The purpose of a CAPTCHA is to separate computer programs from people automatically, using a computer-based test. The typical CAPTCHA user interface consists of two parts: a character image with noise, and an input textbox. The CAPTCHA system will ask the user to type the characters shown in the image. However, the CAPTCHA system will have warped the shape of the characters in the image, and added some arcs or lines to confuse and prevent automated computer recognition of the characters.

*Shih-Yu Huang, Yeuan-Kuen Lee, and Zhan-he Ou are with Department of Computer Science and Information Engineering, Ming Chuan University, Taiwan, R.O.C. (e-mail: syhuang@mcu.edu.tw).

Graeme Bell is with International College, Ming Chuan University, Taiwan, R.O.C.

* To whom all correspondence should be addressed.

Fig. 1 shows an example drawn from the MSN CAPTCHA system. Basically, an automated bot cannot answer the question since until now there is no character recognition technique that can understand what these characters are. On the other hand, humans generally have much better natural abilities when faced with the task of character recognition in a noisy environment, so humans can usually answer these questions correctly without great inconvenience.

Academic research into CAPTCHAs takes the form of a friendly 'arms race', with some researchers acting as 'malicious users' that try to attack and defeat the latest CAPTCHA systems automatically, e.g. [2]-[5], while other researchers seek to design new defensive CAPTCHA techniques in response to known or anticipated attacks [6]-[10]. When designing defensive CAPTCHA techniques, a good CAPTCHA system should give consideration both to computer security and human-friendliness. In practice, balancing these two needs in opposition to one another is very difficult.

In considering the design principles of well-known CAPTCHA systems, we see that many well-known websites such as MSN, Yahoo, Google, Badongo, RapidShare and Youtube are employing user interfaces similar to Fig. 1. Each website employs different heuristics to prevent malicious users. Badongo uses colored lines to clutter the image, and Youtube uses colored blocks; whereas RapidShare uses smaller colored characters as image noise to increase the security. MSN and Yahoo do not use colored character as noise. Instead, they use straight and curved lines as image *clutter* to confuse the defeating program. Although the security of these CAPTCHAs is increased by these heuristics, human-friendliness of them is decreased. Accordingly, many researchers are trying to find other useable principles which will help human users to pass a CAPTCHA test more easily, while still presenting difficulties to automated programs. These principles include techniques such as using high contrast colors, or fewer characters, to assist humans.

Turning now to the task of attacking CAPTCHAs, we observe that this typically involves two procedures, segmentation and recognition. The segmentation procedure requires identification of the correct positions for each character and the recognition procedure identifies which character is in each position. In recent research, [4] shows “segmentation” is a much more difficult problem than “recognition” since machine learning algorithms can efficiently solve the recognition problem, but currently we have no effective general algorithm to solve the segmentation problem causing by these added clutters.

In [5], Chellapilla and other researchers use the image opening and labeling technique to design a segmentation algorithm. When the difference of width between clutters and characters is very noticeable, it is able to separate the noise from characters effectively. However, when the difference is not so noticeable, this algorithm will either be unable to eliminate noise, or it may break the characters when attempting to remove image noise. Therefore, this paper proposes an efficient segmentation algorithm for attacking CAPTCHAs. As mentioned in the previous paragraphs, CAPTCHA systems have a wide variety. It is hard to attack all CAPTCHA tests by a single segmentation algorithm. We assume that the CAPTCHAs being attacked have characteristics as follows: a single-color displayed picture, using warped characters and straight and curved lines as image noisy clutter to confuse the attacking program. The representatives of them are the MSN and Yahoo systems which are used as the basis for the main discussion. This paper introduces a new segmentation technique called projection and outlines a new algorithm, that together make novel and useful contributions to the field of CAPTCHA analysis.

The rest of this paper is organized as follows. Section II illustrates Chellapilla’s segmentation algorithm. Section III presents the proposed projection-based segmentation scheme. Section IV covers experimental results. Section V provides research conclusions.

II. CHELLAPILLA’S ALGORITHM

The algorithm design includes preprocessing, image opening and labeling (three phases) to defeat Yahoo’s CAPTCHA system, shown in Fig. 2. The preprocessing phase includes thresholding and up-sampling – first, converting the original image into a two colored image, and then enlarging it. Image opening is the key step allowing segmentation of the characters from CAPTCHA images. In this phase, the preprocessed image will go through an erosion process several times and will then be dilated several times. Erosion will erase the character borders one pixel per time, whereas dilation will mend the borders one pixel per time. Once the thin clutter items have been deleted by the erosion process, they no longer appear following the dilation process, resulting in some items of clutter being deleted. The labeling phase then finds all of the connected components in the image, and considers the larger ones as characters. Since this phase only outputs the larger items as its result, small connected components will be considered to be clutter and will be eliminated in this phase.

This algorithm is useful when the clutter is of thinner width than the characters. But, this algorithm produces errors because it can not recognize the difference between characters and clutters of similar width. The algorithm may categorize the clutter as character data, so the clutters will not be deleted. An example is shown in Fig. 3(a) and Fig. 3(b), where “S” and “8” are still connected and “G” and “H” have the same problem.

III. PROPOSED SEGMENTATION ALGORITHM

Chellapilla et al. gave the research community an effective way to address the recognition problem, but their segmentation algorithm does not represent a complete solution. This paper will therefore now propose a novel techniques - *projection* - that is intended to improve the success rate of segmentation, and which yield a more effective segmentation algorithm.

A. Projection

The projection technique in this paper is based upon the idea of projecting the image data onto the X-axis. In practice, this is implemented by summing the number of non-white pixels in each column of the image parallel to the Y-axis. Fig. 4 shows some example of clutter and their corresponding projections. It’s easily seen that the projections of these clutter items onto the X-axis appear smaller and flatter than a normal character’s projection onto the X-axis. The projection in the X-axis will tend to appear large and unstable, when a component represents a character rather than an item of clutter. Therefore, by computing a component’s projection value and its variance value it is possible to differentiate between components that are clutter and components that are characters.

Fig. 5(a) gives another type of clutter which is intersected by other characters and forms a part of a larger component. This type of clutter also has a smooth and small appearance when projected into the X-axis, as shown in Fig. 5(b). Therefore, it is possible to use the projection technique to find out the position of the clutter within a component, making it more straightforward to clean up the component. When two or more characters are connected by this form of clutter, they can be effectively split up by deleting these clutter items.

This paper uses a sliding window to this type of clutter, because it has a smaller projection size than a normal character for a small part of the image. In other words, the X-axis projection value of these clutter items will be smaller than some threshold for a part of the image, so it is possible to use the sliding window to check the projection value continually. When the projection values in the sliding window are smaller than the threshold, the algorithm marks the position as containing a clutter item, so that it may be erased.

Fig. 5(b) gives an example of the operation of the sliding window approach with the above clutter. Suppose the width of the sliding window, and the threshold, are both 5. When the sliding window moves to the edge of the "E" and the "5", the projection values in the sliding window are not all smaller than the threshold, so no action will be taken at this position. When the sliding window moves to the edge of the "5" and the "K", all of the X-axis projection values are smaller than the threshold, so the algorithm will mark this place as containing a clutter, and clean it from the image. After the cleaning process, the connection between the "5" and the "K" characters is removed, and these characters are split into separate components.

Note that the projection values of some characters, such as "0", "O", "D", "8", or "B", can also be smaller than the threshold continually over a region of the image, resulting in this character being damaged by the decisions made by the projection method. Fortunately, these characters have closed regions within them, containing the background color. This property can be utilized to avoid damaging the character by mistake.

B. Proposed Algorithm

This algorithm is based on Chellapillas et al.'s algorithm and has five phases, which are preprocessing, opening, labeling, projection and character extracting. The behavior is shown in Fig. 6. The first three phases have a similar process for these of Chellapilla's algorithm except the preprocessing phase. To prevent the mistake generated by the projection technique, the algorithm will detect closed regions in the preprocessing phase. It begins by computing all of the background-colored connected components first. The background-colored connected components which do not belong to a closed region will combine together, and become the largest connected component in the picture. The algorithm marks the connected components that are smaller than the largest component, as closed regions, which will not be considered by the sliding window technique later.

The proposed projection technique is employed in the fourth phase. After operations of this phases, the original image has separated into many different connected components. The final phase, the character extracting phase, deletes the redundant components, and outputs the location of the characters. It is known that the characters have the biggest and most unsmooth projection values in the X-axis, so the algorithm erases the components which have small and smooth projection values. The remaining components are sorted by their size; and the algorithm outputs the largest 8 components (for MSN CAPTCHAs), or the largest 6 components (for Yahoo CAPTCHAs, which typically have 5-6 components).

IV. EXPERIMENTAL RESULTS

In this section experimental results are presented for the Chellapilla's algorithm and the proposed projection-based algorithm. These algorithms are applied to sample images from the MSN and Yahoo CAPTCHA systems. 100 MSN and 100 Yahoo CAPTCHA images are tested for each system. The segmentation rate in these experimental results is based upon the numbers of characters in different images. For example, every image in the MSN system has 8 characters. If the algorithm can segment 20 characters from 10 images, the segmentation rate will be 20/ (10*8) = 0.25, or 25%.

Table 1 gives the benefits of the proposed algorithm. The segmentation rate of the proposed algorithm attacking the Yahoo CAPTCHA system was higher than Chellapilla's algorithm by 9%. Attacking the MSN CAPTCHA system, the algorithm can also yields a 14% higher segmentation rate. Fig. 7 and Fig. 8 show a set of examples for both Chellapilla's algorithm and the proposed algorithm. In the Yahoo CAPTCHA, Chellapilla's algorithm leaves a component which is made from the clutter, but the proposed algorithm can totally separate all the characters without leaving any clutter. In the MSN CAPTCHA in this example, Chellapilla's algorithm cannot split the "5" and "K", and leaves many items of clutter, but the proposed algorithm can separate the "5" and "K" and can also remove most of the clutter items.

V. CONCLUSIONS

In this paper, an effective algorithm using a new 'projection' technique was proposed for segmentation of the Yahoo and MSN CAPTCHA systems. In the experimental results, it was found that the proposed algorithm can uniformly improve the segmentation rate over the traditional algorithm. The proposed algorithm makes novel and useful contributions to the field of CAPTCHA analysis.

REFERENCES

[1] C. Pope and K. Kaur, "Is It Human or Computer? Defending E-Commerce with Captchas," IT Professional, vol. 7, no. 2, pp. 43–49, 2005.

[2] G. Mori and J. Malik, "Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA," in Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 134-141, 2003.

[3] G. Moy, N. Jones, C. Harkless, and R. Potter, "Distortion Estimation Techniques in Solving Visual CAPTCHAs," in Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 23-28, 2004.

[4] K. Chellapilla, K. Larson, P. Simard, and M. Czerwinski, "Computers Beat Humans at Single Character Recognition in Reading Based Human Interaction Proofs (HIPs)," in

Proceedings of the Third Conference on E-Mail and Anti-Spam, 2005.

[5] K. Chellapilla and P. Simard, “Using Machine Learning to Break Visual Human Interaction Proofs (HIPs),” in L. K. Saul, Y. Weiss, and L. Bottou, editors, Advances in Neural Information Processing Systems 17, pp. 265–272. MIT Press, Cambridge, MA, 2005.

[6] L. Coates, H. S. Baird, and R. J. Fateman, "Pessimal Print: A Reverse Turing Test," in Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp. 1154-1158, 2001.

[7] M. E. Hoque, D. J. Russomanno, and M. Yeasin “2D Captchas from 3D Models,” in Proceedings of the IEEE SoutheastCon, pp. 165-170, 2006.

[8] H. S. Baird and J. L. Bentley, “Implicit CAPTCHAs,” in Proceedings of Document Recognition and Retrieval XII, pp. 191-196, 2005.

[9] M. Shirali-Shahreza and S. Shirali-Shahreza, “Drawing CAPTCHA,” in Proceedings of the 28th International Conference on Information Technology Interfaces, pp. 475-480, 2006.

[10] D. Misra and K. Gaj, “Face Recognition CAPTCHAs,” in Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services, pp. 122-127, 2006.

Table 1: Experimental results.

	Yahoo (total characters: 525)		MSN (total characters: 800)	
	Correct number	Segmentation rate	Correct number	Segmentation rate
Chellapilla's algorithm	318	60.57%	329	41.13%
Proposed algorithm	367	69.90%	441	55.13%

Fig. 1. An example of the MSN CAPTCHA system.

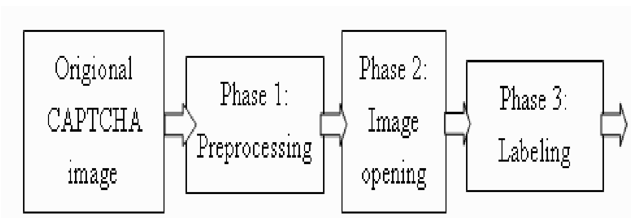


Fig. 2. Structure figure for Chellapilla's CAPTCHA segmentation algorithm.

(a) (b)

Fig. 3. Some problems in Chellapilla's algorithm.
(a) Original image. (b) Some clutter can not be erased.

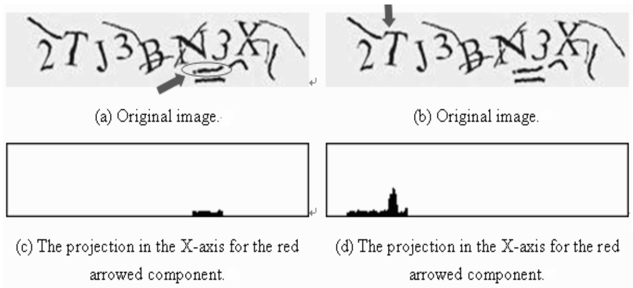


Fig 4. Examples of clutter and its projection image.

(a) (b)

Fig. 5. Example of clutter intersected by other characters and its projection image.

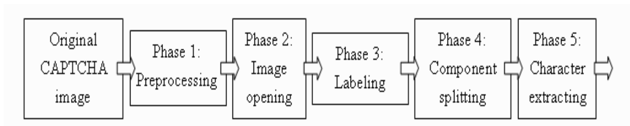
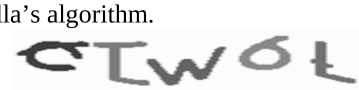


Fig. 6. Structure diagram for the proposed CAPTCHA segmentation algorithm.

(a) Original image.

(b) Chellapilla's algorithm.

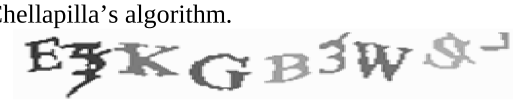


(c) Proposed algorithm.

Fig. 7. Algorithm results of Yahoo CAPTCHA system.

(a) Original image.

(b) Chellapilla's algorithm.



(c) Proposed algorithm.

Fig.8. Algorithm results of MSN CAPTCHA system.